Charlie Carpenter

BIOS 7659

Final Project Report

**Introduction:**

This report is based on a 2015 study conducted by Kiselev, et. al, *Perturbations of PIP3 signalling trigger a global remodelling of mRNA landscape and reveal a transcriptional feedback loop*. The dataset their paper is based on is housed in the Gene of Expression (GEO) repository under the accession number GSE69822. In this paper they establish several feedback loops in the mRNA of MCFA10a cells caused by acute perturbations of Class I phosphoinositide-3-kinases (PIP3) signaling under 5 different experimental conditions. This report will only focus on the comparison of two of these conditions in order to apply a different analysis technique for this time course design.

PI3Ks are lipid kinases that are important operators for signal transduction events in cells. These signal transduction evens allow the cells’ surfaces to control key cellular functions, such as growth, movement, proliferation, and differentiation. Class I PI2Ks will catalyze the phosphorylation of phosphatidylinositol-4,5-bisphosphate (PI(4,5)P2) to produce phosphatidylinositol-3,4,5-trisphosphate (PIP3) in the plasma membrane once the necessary cell surface receptors are activated1. Other lipid phosphates control PIP3 levels throughout the degradation process; PTEN is a compound that disrupts this interaction with the other lipid phosphates.

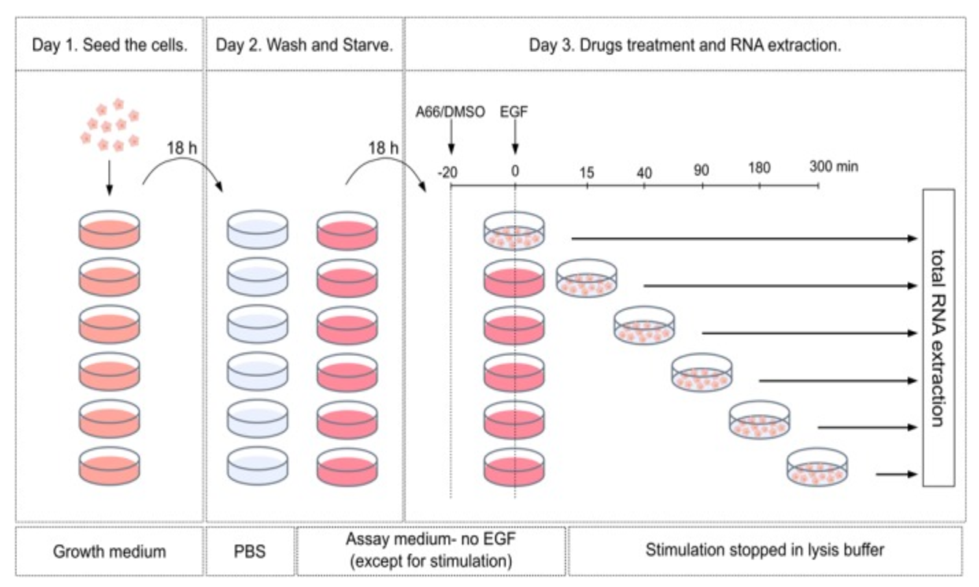
Strong research2-3 has shown evidence that the initial impacts of activating PI3K pathways propagate by allosteric and post-translational effects, which in turn impacts the regulation of translation, mRNA stability, and regulation of transcriptome factors. This research has also identified homeostatic feedback loops and evidence of regulation of critical features of the pathway by mRNA networks (e.g. mRNA regulation of PTEN). These features are essential for the integrity of the pathway of PI3K activation. In this report we explore the impacts of PTEN perturbation on mRNA accumulation in human breast epithelial MCF10a cells.

The original analysis included pairwise comparisons of each condition at each time point and a comparison of each condition over multiple time points, but it primarily focused on searching for significantly enriched GO terms. Both the pairwise and time course analyses were accomplished using DESeq2 from Bioconductor5. This report uses impulse models6, which have been shown to have higher statistical testing power over DESeq2 in time course experiments with many (≥ 6) time points7, to assess the differences in trajectories in gene expression between the wild type and PTEN knockout cells. The original ImpulseDE6 can be applied to any high throughput gene expression data, but the algorithm was streamlined for count data in the ImpulseDE2 package in Bioconductor7. The package offers the ability for case-control time course experiments along with several other handy features such as batch effect corrections. These will not be explored as they are not relevant to this data set.

**Methods:**

Experimental Design:

The non-transformed human breast epithelial MCF10a cells were obtained from Horizon Discovery. The starvation assay and RNA-seq assays were made of DMEM/F12 supplemented with 1% charcoal dextran treated fetal bovine serum (FBS), 0.1 μg/ml cholera toxin, 0.2 μg/ml hydrocortisone. These compounds came from Life Technologies and PAA.

Day 1 of cell treatment involved seeding 320,000 cells in 60mm dishes with a growth medium. These were left to adhere to the dishes for 18 hours at 37°C. On day 2, the cells are washed with phosphate buffered saline, and then the growth medium is replaced with the assay medium. Cells were then starved in a growth serum, EGF, insulin for 18 hours at 37°C. On day 3, cells were incubated with DMSO for 20 minutes, then incubated with EGF (10ng/ml) for a further 15, 40, 90, 180 or 300 minutes (Figure 1). This process was repeated 3 times for all 5 of the original experimental   *Figure 1: Illustration of the experimental procedure*

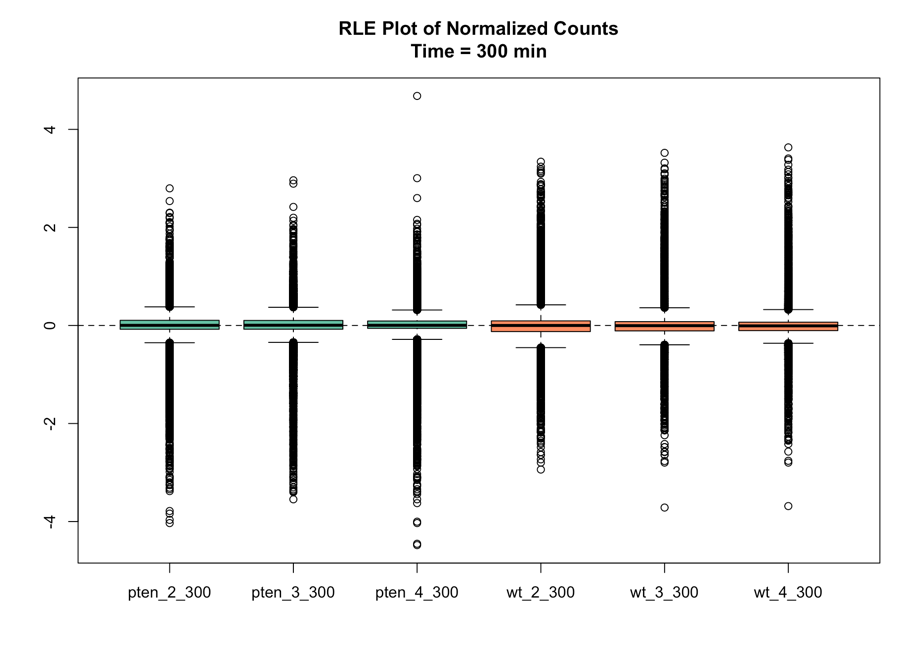
conditions. We will only focus on the PTEN knockout (KO) and wild type (WT) at each of the 6 time points to illustrate the ImpulseDE2 model.

The RNA-seq libraries were prepared using the Truseq RNA sample preparation kit from Illumina following the protocols recommended by Illumina. Pools of six libraries were prepared for 100-bp single end sequencing on a HiSeq 2500 machine from Illumina.

Statistical Analysis:

*Quality Control*

Fastq files were not provided to the public through the GEO dataset for the original authors’ data. For their publication, they performed FastQC4 state that “About 85–95% of reads in every sample were uniquely mapped, with the majority of samples having >90% uniquely mapped reads” based on the “human reference genome (Ensembl GRCh37 release 75).” Analysis was performed assuming that this quality of reads still held for our subset of experimental conditions.

Genes with less than 5 reads over all time points in each replicate in either the PTEN KO or WT genes were removed from analysis. This gave us 18,739 genes to analyze. Samples were normalized using full non-linear *Figure 2: Normalized counts at Time = 300 minutes* quantile normalization using the EDASeq package from Bioconductor9. Figure 2 shows the normalized counts for the PTEN KO and WT genes at their final time point. This plot is representative of what we see from the other 5 time points. None of the technical replicates stood out from the others, so all will be included for all 6 time points.

*Differential Expression*

Impulse Models

Impulse models are designed to capture the initial, peak, and steady state of expression after some experimental perturbation. They are defined as the product of two separate sigmoid models and scaled by the peak state of expression. The two separate sigmoid models represent the shifts from initial to peak state and peak to steady stat expression. This can be expressed as

where , , and represent the initial, peak, and steady state levels of expression, respectively, and and are the transition times to peak and steady state, respectively. Note that this can collapse into a single sigmoid model if for monotonous expression growth when . The option to fit sigmoid models to each gene is offered by the ImpulseDE2 package if the investigator believes that a steady state of expression is not reached by the end of the trial.

Unlike ImpulseDE, ImpulseDE2 is designed to handle count data such as transcriptomic read counts. It assumes that our observed number of reads from µ transcripts comes from a negative binomial distribution like other popular methods such as DESeq2 and edgeR. The negative binomial likelihood of the data for every gene from samples at time points is defined as

where is a gene and time (and condition) specific mean parameter and is a gene specific dispersion parameter. The dispersion parameter is considered a hyper-parameter to be estimated for each gene before is calculated. ImpulseDE2 estimates using DESeq2 treating time as a categorical variable. is a normalization factor that is based on an estimated *size* factor *s* and *batch* factor *b* for the batch such that

Here the size factor is the median ratio of the observed counts to their geometric mean per gene.

Algorithm

The primary function for ImpulseDE2 requires a matrix of raw counts along with a additional table of meta data. The algorithm follows 3 main steps:

1. **Hyper-parameter estimation:** DESeq2 is run for each gene and dispersion parameters are stored for the impulse models.
2. **Model Fitting:** For each gene
   1. Two impulse models are fit based the highest and lowest value of expression to provide initial estimates.
   2. Optimized impulse parameters based on the negative binomial log likelihood are achieved through the Broyden-Fletcher-Goldfarb-Shanno algorithm.
   3. When in a case-control setting this is done for the case samples, control samples, and the “complete” data set.
3. **Differential expression:** For each gene a loglikelihood ratio test is performed and p-values are obtained from a distribution. The p-values are then corrected with the Benjamini-Hochberg false-discovery rate.

**Results:**

For this report we consider FDR corrected p-values below 0.0001 are considered to be significantly different. Of the 18,739 genes to analyzed, there were 5624 genes with strong evidence of different expression trajectories between the PTEN KO and WT genes. The top 15 genes based on the FDR adjusted p-value.

*Table 1: Subset of genes with significantly different expression trajectories*

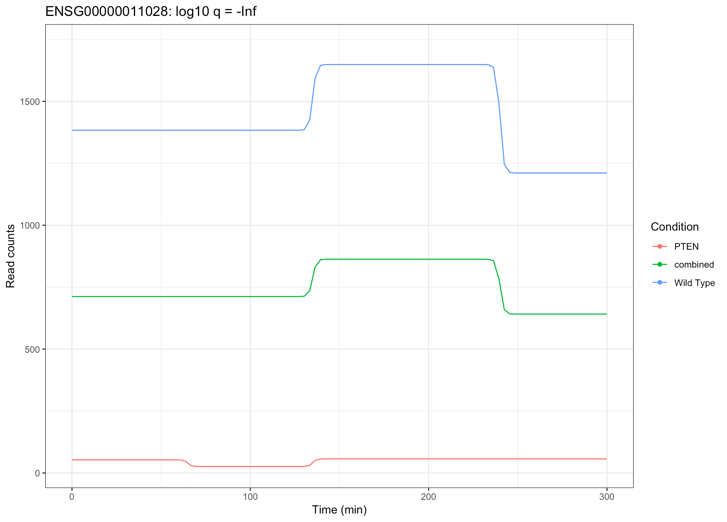
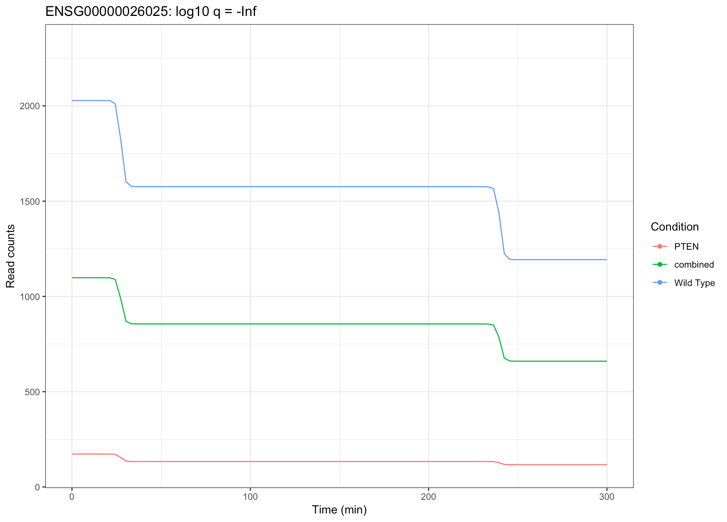
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Gene | p | FDR  p-value | loglik\_full | loglik\_red | df\_full | df\_red |
| ENSG00000002746 | <0.0001 | <0.0001 | -111.79054 | -1044.9631 | 13 | 7 |
| ENSG00000003436 | <0.0001 | <0.0001 | -218.48421 | -1032.9923 | 13 | 7 |
| ENSG00000008838 | <0.0001 | <0.0001 | -244.61721 | -2382.1205 | 13 | 7 |
| ENSG00000011028 | <0.0001 | <0.0001 | -175.74094 | -2912.8379 | 13 | 7 |
| ENSG00000016391 | <0.0001 | <0.0001 | -102.72703 | -1402.2684 | 13 | 7 |
| ENSG00000023839 | <0.0001 | <0.0001 | -238.46701 | -2743.2108 | 13 | 7 |
| ENSG00000026025 | <0.0001 | <0.0001 | -205.52560 | -1222.7648 | 13 | 7 |
| ENSG00000054219 | <0.0001 | <0.0001 | -97.13328 | -1144.7278 | 13 | 7 |
| ENSG00000063601 | <0.0001 | <0.0001 | -195.24650 | -1163.3295 | 13 | 7 |
| ENSG00000064042 | <0.0001 | <0.0001 | -210.82074 | -1589.9976 | 13 | 7 |
| ENSG00000071246 | <0.0001 | <0.0001 | -117.06268 | -957.4972 | 13 | 7 |
| ENSG00000073008 | <0.0001 | <0.0001 | -236.47595 | -1437.6475 | 13 | 7 |
| ENSG00000073849 | <0.0001 | <0.0001 | -188.22059 | -2921.6155 | 13 | 7 |
| ENSG00000076344 | <0.0001 | <0.0001 | -119.05949 | -1531.9207 | 13 | 7 |
| ENSG00000076706 | <0.0001 | <0.0001 | -119.63017 | -2380.9037 | 13 | 7 |

**Discussion:**

The PTEN compound has large impacts on the genetic expression of human breast epithelial MCF10a cells. Unfortunately, these results are not directly comparable to the original analysis since the original focused heavily exploring clusters of expressed genes through GO analysis. The impulse model should be an improvement over the DESeq2 approach for time course models since it treats time as a continuous variable. The method easily implemented through the ImpulseDE2 package from Bioconductor.

One improvement that could be made is adding some diagnostic or tool to aid in the decision to fit either a sigmoid or impulse model. The package allows for both to be implemented easily, and the impulse model is not necessary when we have monotonous expression change. Figures 3 shows the expression trajectories for two different genes where the impulse model seems more appropriate for the first, but a sigmoid model should suffice for the second. We wouldn’t know this unless we dug around through each gene and specified this ourselves.

That said, the impulse model is also unintentionally appropriate for modeling expressions that are expected to experience “feedback” loops like we have in this data set. The feedback loop could create two separate shifts in expression in the same direction like we see in the gene ENSG00000026025 (right) in Figure 3. This could be an improvement on the sigmoid model since we are capturing the monotonous growth and the “feedback” aspect simultaneously.

*Figure 3: The trajectory of the gene ENSG00000011028 (left) would require an impulse model to detect the peak and steady states in its trajectory, but the gene ENSG00000026025 (right) has a monotonous trajectory that might be better fit by a sigmoid model.*

**Reference:**

1. Vanhaesebroeck B., Guillermet-Guibert J., Graupera M., Bilanges B. The emerging mechanisms of isoform-specific PI3K signalling. Nat. Rev. Mol. Cell Biol. 2010;11:329–341.
2. Manning B.D., Cantley L.C. AKT/PKB signaling: navigating downstream. Cell. 2007;129:1261–1274.
3. Schmidlin M., Lu M., Leuenberger S.A., Stoecklin G., Mallaun M., Gross B., Gherzi R., Hess D., Hemmings B.A., Moroni C. The ARE-dependent mRNA-destabilizing activity of BRF1 is regulated by protein kinase B. EMBO J. 2004;23:4760–4769.
4. Andrews S. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> 2010
5. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550–571.
6. Jil Sander, Joachim L Schultze, Nir Yosef; ImpulseDE: detection of differentially expressed genes in time series data using impulse models, Bioinformatics, Volume 33, Issue 5, 1 March 2017, Pages 757–759, <https://doi.org/10.1093/bioinformatics/btw665>.
7. David S. Fischer, Fabian J. Theisa, and Nir Yosef; Impulse model-based di\_erential expression analysis of time course sequencing data, March 3, 2017.
8. David S Fischer (2018). ImpulseDE2: Differential expression analysis of longitudinal count data sets. R package version 1.6.0.
9. D. Risso, K. Schwartz, G. Sherlock and S. Dudoit (2011). GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 12:480